唇領域の動静判定と音声・雑音判定の統合に基づく発話区間の検出

増田　健† 松田　博義† 井上　淳一† 有木　康雄†† 滝口　哲也††

古賀健太郎†††

† 神戸大学大学院自然科学研究科
兵庫県神戸市灘区六甲台町 1-1
†† 神戸大学工学部
兵庫県神戸市灘区六甲台町 1-1
††† 富士通テン株式会社
兵庫県神戸市兵庫区御所通 1-2-28
E-mail: †{masudaken,matsuda,inoue}@me.cs.scitec.kobe-u.ac.jp, ††{ariki,takigu}@kobe-u.ac.jp,
†††k-koga@mms.ten.fujitsu.com

あらまし　音響信号と同時に目的発話者の顔画像を利用し，唇領域の情報を抽出することで発話区間検出を行う手法を提案する．周囲からの雑音，特に目的話者の近くで発話している人の音声区間を誤って検出するという問題を防ぐため，顔画像より抽出した唇領域を用いて動静判定を行い，発話の有無を検出する．また，画像の撮影には，照明条件の変化にも対応できるように赤外線カメラを用いている．このため，グレースケール画像より抽出できる特徴量として一般的である明度値をもとに，目的領域の探索を有効に行う手法として，Haar 状特徴を用いた AdaBoost 法を採用している．更に，正規化相関法を用いた唇領域の追跡，抽出を行うことで，画像から取得できる情報量を増やして，検出精度を向上している．実際に車内で撮影されたデータを使った実験より，提案手法を用いることで，適合率は音声判定のみの結果と比較して平均で 25% 向上した．

キーワード　発話区間検出，唇領域抽出，正規化相関法，AdaBoost，GMM

Voice Activity Detection by Integrating Lip Open/Close Discrimination and Speech/Noise Discrimination

Ken MASUDA†, Hiroyoshi MATSUDA†, Junichi INOUE†, Yasuo ARIKI††,

Tetsuya TAKIGUCHI††, and Kentarou KOGA†††

† Graduate School of Science and Technology , Kobe University
1-1, Rokkodai , Nada , Kobe , Hyogo
†† Department of Computer and Systems Engineering , Kobe University
1-1, Rokkodai , Nada , Kobe , Hyogo
††† FUJITSU TEN Corporation
1-2-28, gosyodori , hyogo , kobe , hyogo
E-mail: †{masudaken,matsuda,inoue}@me.cs.scitec.kobe-u.ac.jp, ††{ariki,takigu}@kobe-u.ac.jp,
†††k-koga@mms.ten.fujitsu.com

Abstract　We propose a voice activity detection by using lip area movement and sound signals. To prevent the wrong detection caused by the different speakers, our system watches the lip movement of the target speaker. We use an infrared camera to cope with the change of lighting environment. We adopt the method of AdaBoost based on the Haar like feature to extract the lip area effectively. Haar like feature can be easily computed from gray scale image. The proposed system improved the precision rate by 25% than that achieved by using sound signals recorded in a car.

Key words　voice activity detection , lip area extraction , regularized correlation , AdaBoost , GMM

1. はじめに

音声による意思伝達手法は，ボタン操作などによる入力時の煩わしさがなく，常に両手が塞がっている状態である運転中のドライバーなどにとって役に立つ．しかし，車内のように狭く雑音の多い環境下で，音響信号のみを用いてドライバーの要求を認識しようとすると，雑音や助手席の音声など目的話者以外の音響信号によって誤認識の可能性が生じる．

この問題を解決するために，雑音を判別して除去する方法が考案されているが，検出された音声がドライバーによる音声なのか，助手席を含む他の人間による音声なのかを，音響信号のみで判別することは困難である．そこで，本論文では，この音声区間からドライバーの発話区間のみを検出することを目的として，唇領域の抽出・追跡と動静判定処理を用いたシステムの構築を提案する．

音響信号と同時にドライバーの顔を含む画像情報を用いることで，目的話者の発話区間を検出しようとする研究は盛んに行われている [1] [2] [3] が，その多くは，音声と深く関わる唇領域の抽出と，その形状の把握による発話の有無を調べるというものである．唇領域を抽出するにあたって，RGB 値分布 [4] [5]，HSV 値分布 [6]，エッジ抽出 [6] [7]，動きベクトル [8] など様々な特徴が用いられている．また，目的領域を抽出する手法として，テンプレートマッチングによる抽出 [9] [10]，SNAKE [7] [11]，ヒストグラム比較によるアクティブ探索法 [12] [13] [14]，Boosting 法 [15] などはよく知られている．

しかし，カラー情報の利用は照明環境に左右されやすいため，夜間走行などが考えられる車内で使用するには適さず，また実時間処理という視点より計算コストは低い方が望ましい．従って，提案手法では，夜間でも使用可能な赤外線画像を用いる．赤外線画像はグレースケール画像であり，特徴量として得られる明度値から高速に唇領域を抽出可能な手法として AdaBoost 法を採用している．また，唇領域の追跡手法として照明変化に頑健な正規化相関法を用いる．さらに，発話に伴って唇領域における明度値分布の変化量が大きくなることを考慮し，動静判定を行うことで口が動いている区間を判定する．最後に，音響信号を基に音声と雑音を判別して雑音区間を削除した音声区間を検出し，その結果と統合することで，ドライバーの発話区間のみを検出する．

本論文の構成について述べる．

2 節では，画像情報を用いた唇領域の動静判定のアルゴリズムについて述べる．3 節では，GMM を用いた音声区間検出のアルゴリズムについて述べる．4 節では，画像情報と音響信号を用いた発話区間の検出結果の統合について述べる．5 節で，提案手法を用いた評価実験とその結果について述べる．

2. AdaBoost と正規化相関法を用いた唇領域の動き検出

本節では，画像から唇領域を抽出・追跡し，更にフレーム間での領域差分の和を求めることにより唇の動静を判定する方法について述べる．追跡には正規化相関法によるテンプレートマッチングを適用し，テンプレートの初期値を得るために Haar 状特徴を用いた AdaBoost 法による唇領域の抽出を行う．以下に，その流れを述べる．

2.1 AdaBoost による初期位置の決定

上述したように，単純に Haar 状特徴を用いた AdaBoost 法を適用しただけでは，唇領域に類似した領域も唇領域の候補として抽出してしまう可能性がある．従ってここでは，探索領域において更に唇が存在する可能性の高い領域を絞り込むことによって，真に唇領域である候補を唯一決定する方法を提案する．

ドライバーの顔画像を含む動画の初期 30 フレームを用いて，次の方法により唇領域を確定した．なお，ここで使用した動画は全て 30 フレーム毎秒である．

（1） 顔画像を N ブロックに分割する（図 1-(b)）．

（2） 各フレームで抽出された唇候補領域の重心を求め，それぞれの重心が属するブロックに 1 点を投票していく（図 1-(c), (d)）．候補は複数存在することがあるため，1 つのブロックが 1 点以上得票することがある．30 フレームが終了した時点で得票数が最大のブロックを求める（図 1-(e)）．

（3） 31 フレーム目において，上述の方法により求めたブロック内に重心が存在する唇領域候補で，最も中央に位置する候補をテンプレート用の唇領域として選出する（図 1-(f)）．

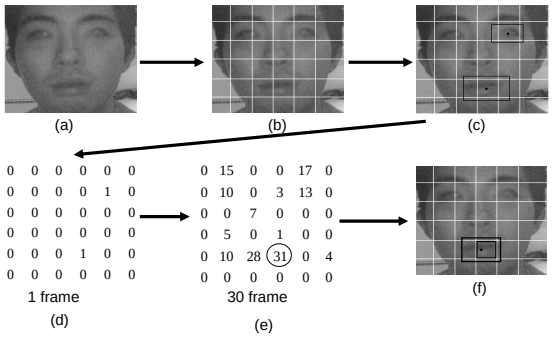

図 1　唇領域の決定

2.2 正規化相関法によるテンプレートマッチング

（1） 直前フレームで抽出された唇領域をテンプレート（図 2-（Lip_{temp}））として，現在のフレームにおけるサイズ (W, H) の探索領域（図 2-（Search area））に適用する．

（2） 探索領域内の座標 (a, b) で，式（1）を用いて，正規化相関値 $V(a, b)$ を求める．ただし，テンプレート内の座標 (i, j) における値を $t(i, j)$，探索領域内の座標 (i, j) における値を $s(i, j)$ とし，

$$\bar{t} = \frac{\sum_{i \leqq w} \sum_{j \leqq h} t(i,j)}{h \cdot w}$$

$$\bar{s} = \frac{\sum_{i \leqq w} \sum_{j \leqq h} s(a+i, b+j)}{h \cdot w}$$

とする．

$$V(a,b) = \frac{\sum_i \sum_j \{t(i,j) - \bar{t}\}\{s(a+i,b+j) - \bar{s}\}}{\sqrt{\sum_i \sum_j \{t(i,j) - \bar{t}\}^2} \sqrt{\sum_i \sum_j \{s(a+i,b+j) - \bar{s}\}^2}} \quad (1)$$

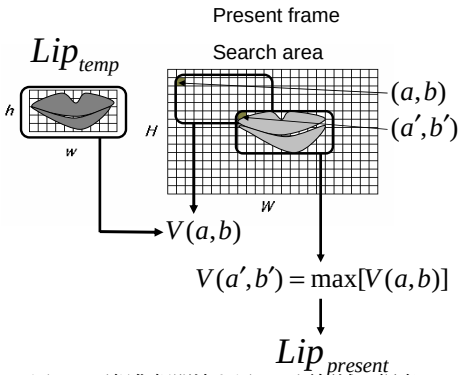

図 2　正規化相関法を用いた唇領域の探索

$$(0 \leqq a \leqq W - w, \quad 0 \leqq b \leqq H - h)$$

（3） $V(a,b)$ が最大値をとる探索範囲内の座標 (a',b') を始点とする領域 (w,h) を現在のフレームにおける唇領域（図 2-($Lip_{present}$)）とする．

（4） 探索窓の更新を行う．

$$Lip_{temp} \leftarrow Lip_{present} \tag{2}$$

2.3　唇領域の絶対値差分和による動静判定

式（3）より，唇の参照領域 Lip_{sub} と，現在のフレームで抽出された唇領域 $Lip_{present}$ の全画素値に対する絶対値差分和を求める．今回は，この参照領域 Lip_{sub} として，直前のフレームより抽出された領域 $Lip_{sub}(x,y) = Lip_{temp}(x,y)$ を採用する．

式 (3) の S が閾値 ζ 以上であれば動きがあると判定し，未満であれば動いていないと判定する．

$$S = \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} |Lip_{present}(x,y) - Lip_{sub}(x,y)| \tag{3}$$

$$S \geqq \zeta \rightarrow Move$$
$$S < \zeta \rightarrow Stop$$

3.　GMM による音声区間検出

3.1　音声特徴抽出

音声特徴抽出には MFCC(Mel Frequency Cepstrum Coefficients) を用いる．MFCC は各時刻におけるメル周波数スペクトルを DCT により直交変換（ケプストラム化）した静的な特徴である．時間的に動的な特徴も含めるため，各フレームから得られた MFCC 音声特徴ベクトルを，前後の 2 フレームとあわせる．それにより得られる計 5 フレームからなるベクトルを，一つの音声特徴ベクトルとして扱う．ただしこれらの手順は，5 フレームごとにブロック単位で行うのではなく，1 フレームごとに窓をシフトさせることによって行う．この手法により，尤度計算の際に計算コストが大きくなるが，雑音と音声をより明確に判別することが可能になる．[16]

3.2　尤 度 計 算

前述した音声特徴に対して，GMM(Gaussian Mixture Model:混合正規分布) を用いて尤度計算を行う．GMM とは，式（4）で表されるように，複数の正規分布の重みつき和をとることにより，複数の山を持たせた分布のことであり，次式で定義される．

$$f(o) = \sum_{w=1}^{W} \lambda_w N(o; \mu_w, \Sigma_w) \tag{4}$$

$$\sum_{w=1}^{W} \lambda_w = 1 \tag{5}$$

ここで、D はデータの次元数，W は混合数，$o \in R^D$ はデータのベクトル，$\mu_\omega \in R^D$ はω番目の分布に属する平均ベクトル，$\Sigma_\omega \in R^{D \times D}$ は分散共分散行列であり，$N(o; \mu_\omega, \Sigma_\omega)$ は正規分布である．

3.3　音声区間検出

音声の各フレームから得られた尤度値に対し，式 (6) により，尤度比検定を行う．

$$L(x) = \log \frac{P(x \mid speech)}{P(x \mid noise)} \tag{6}$$

ここで，$P(x \mid speech)$ は音声尤度，$P(x \mid noise)$ は非音声尤度である．$L(x)$ の値は，以下のように前後のフレーム間で平滑化をしておく．

$$L(x_i) = \sum_{j=i-\frac{N}{2}}^{i+\frac{N}{2}} L(x_j) \tag{7}$$

こうして得られた $L(x)$ の値が，ある閾値 θ 以上であれば音声，以下であれば非音声フレームとする．さらに，得られた区間のうち，短い区間を削除することによって，最終的な音声，非音声区間を得る．

4.　音響情報と画像情報の統合

本節では，音響信号より検出した音声区間と，2 節で述べた唇の動静判定の手法を用いて，実際に発話区間を検出する手法について述べる．図 3 に，システム全体の処理の流れを示す．

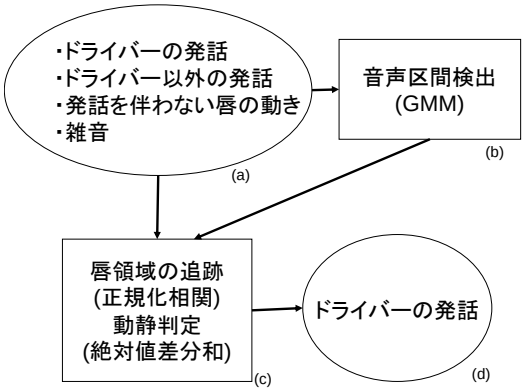

図 3　統合システム

音響信号より検出された音声区間内（図 3-(b) ）でのみ，動静判定（図 3-(c) ）を行うこととする． 各音声区間内で行われ

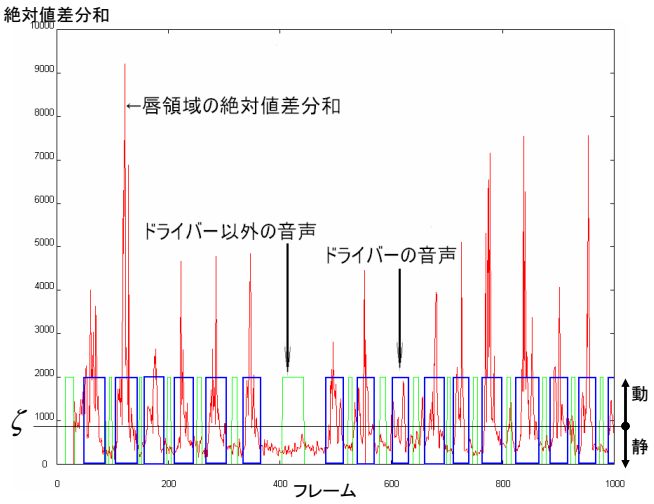

図 4　絶対値差分和を用いたドライバーの音声区間検出

た絶対値差分和による動静判定の結果，唇が動いていると判定されたフレームが存在する音声区間をドライバーによる発話区間，1 度も唇が動かなかったと判定された音声区間を非発話区間とした（図 3-（d））．図 4 に，縦軸を絶対値差分和，横軸をフレーム番号としたときの判定の例を示す．また，閾値 ζ は，適合率が最大になるように設定し，各フレームにおいて出力される絶対値差分和の値が ζ 以上であれば唇は動いていると判定し，ζ 未満であれば唇は動いていないと判定する．

5. 評 価 実 験

提案した手法を用いて，その有効性を調べる評価実験を行った．

5.1 実 験 条 件

学習用に約 7000 枚の唇部分を切り出した画像を用意し，2 節で述べた AdaBoost 法を用いて学習を行った．AdaBoost 法によって決定された強判別器の最適カスケード結合数は 14 であった．図 5 に，学習画像の例を示す．

図 5　学 習 画 像

提案手法を用いて実験を行うにあたり，テストデータとして日本人男性 3 名と女性 2 名の顔画像を含む動画を使用した．1 人当たり昼間と夜間の 2 回，アイドリング状態の車内にて，日本全国の地名 100 単語を発声してもらい，撮影と録音を行った．音響信号におけるノイズの SN 比はカットオフ周波数 200Hz のハイパスフィルタをかけて，10dB〜20dB であった．既に述べたように，夜間画像にも対応可能な赤外線カメラを用いている．また，動画を使用する際には，ファイルサイズの問題から DivX-Codec [17] を用いて圧縮した．図 6 に，テストデータの例を示す．上段が昼間に撮影した画像で，下段が夜間に撮影した画像である．

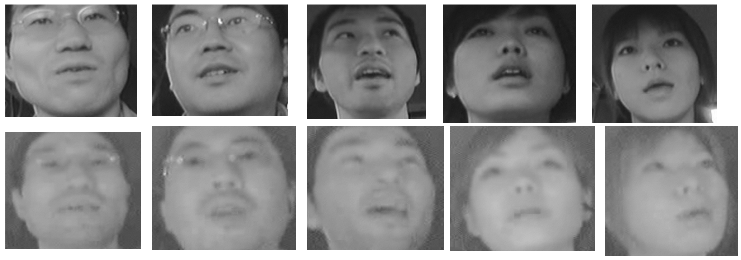

図 6　テストデータ

表 1　正解率 95% 以上

Speaker ID	Detect all	Detect true	Recall	Precision
A:day time	124	96	96	77.42
B:day time	104	97	97	93.27
C:day time	113	97	97	85.84
Dday time	108	100	100	92.59
E:day time	120	96	96	80
Mean：day time	113.8	97.2	97.2	85.41
A:night	119	96	96	80.67
B:night	113	96	96	84.96
C:night	122	96	96	78.96
D:night	126	96	96	76.19
E:night	135	97	97	71.85
Mean：night	123	96.2	96.2	78.21
Mean：all	118.4	96.7	96.7	81.67

動画の撮影は，図 7 に示すように，車内のドライバー席の正面位置に取り付けられたカメラを用いて行った．

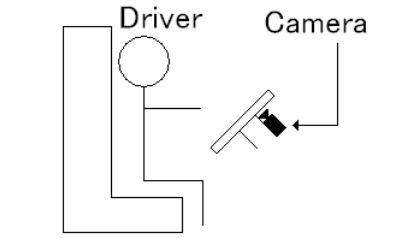

図 7　撮 影 環 境

録音内容より，GMM を用いて音声区間を検出する．今回の実験で用いたデータの収録時は，車内にドライバーしか存在していなかったため，周囲環境からの雑音はあってもドライバー以外の人間の発話区間が誤検出されることはなかった．従って，検出された各単語の発話区間の間に，ドライバー以外の発話を擬似的に挿入し，ドライバーの発話区間検出実験を行った．図 8 に，音響信号より検出した音声区間結果の例を示す．横軸はフレーム番号である．図の上段（a）はドライバーのみが発声したデータに対して，手動で切り出した正解音声区間である．下段（b）はドライバーの発声の間にドライバー以外の発声を挿入したデータに対して，3 節の音声処理によって音声区間を検出した結果の例である．この（b）に対して，画像処理による判定を重ねた結果が（a）に近づくほど，音声区間検出の精度が高いと言える．

なお，音声区間のみの検出結果は，全音声区間検出数（Detect All）が 200，ドライバーの真の発話区間の検出数（Detect True）は 100 であったことから，正解率（Recall）は 100% ，

表 2　正解率 100% 以上

Speaker ID	Detect all	Detect true	Recall	Precision
A:day time	138	100	100	72.46
B:day time	123	100	100	81.30
C:day time	122	100	100	81.97
D:day time	108	100	100	92.59
E:day time	127	100	100	78.74
Mean：day time	123.6	100	100	80.91
A:night	152	100	100	65.79
B:night	120	100	100	83.33
C:night	129	100	100	77.52
D:night	157	100	100	63.69
E:night	156	100	100	64.10
Mean：night	142.8	100	100	70.03
Mean：all	133.2	100	100	75.08

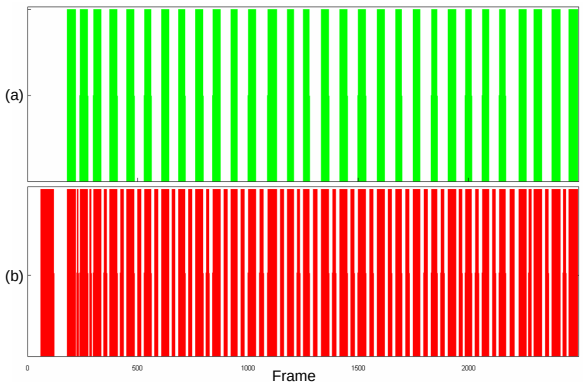

図 8　音声区間検出結果
(a)correct voice section
(b)experimental result

適合率（Precision）は 50% であった．

テストデータにおいて，実際にドライバーが発話した単語の数は 100 であるため，各単語の間と先頭に，擬似的に挿入したドライバー以外の音声区間を含めた全音声区間は 200，検出されるべき正しい発話区間は今回の実験では常に 100 となる．

正解率，適合率は以下の式で求めた．

$$\text{正解率} = \frac{Detect\ \ True}{100} \times 100(\%)$$
$$\text{適合率} = \frac{Detect\ \ True}{DetectAll} \times 100(\%)$$

5.2　絶対値差分和を用いた動静判定結果

絶対値差分和を用いて，唇領域の動静判定を行った結果を示す．直前フレームより抽出した領域との絶対値差分和を求めた場合，正解率を 95% 以上に指定した結果を表 1 に，100% に保った結果を表 2 に示す．表に示した結果より，正解率を損なわないようにした場合，適合率が 50% から約 75% まで向上した．また，正解率を 95% 以上に保ちつつ適合率を最大まで上げるという条件下で閾値を設定した場合，更に 5% ほど向上した．

ドライバーの音声区間検出を行う際，図 4 における閾値 ξ を高く設定するほど唇領域が大きく動いている発話区間のみを検出することになり，雑音などによる唇領域の絶対値差分和の変化検出を軽減することができる．しかし，この閾値を高くしすぎた場合，唇の動きが少ない正解発話区間は雑音と見なされて検出されなくなる可能性が高くなる．実際に閾値を高くした場合，唇領域の変化が小さい正解発話区間は検出されず，変化が大きく確実にドライバーが発話を行ったと思われる発話区間のみを検出したため，正解率は約 95% まで低下したものの適合率は向上した．以上のことから，今後は最適な閾値を求めることが重要となってくる．

また，差分和を用いた場合，昼間よりも夜間の方が適合率の向上は少なかった．夜間に撮影した画像は昼間に撮影した画像と比較して，照明環境の悪化からグレースケール画像の階調性が低くなったため，唇領域の検出精度とそれに伴う絶対値差分和を用いた唇の動き検出精度が低下したためだと考えられる．

6.　ま　と　め

本研究では，車内という限定された環境下において，赤外線画像と音響信号を用いることにより，周囲からの雑音とドライバー以外の音声を除去し，効果的にドライバーの音声のみを検出するシステムの提案を行った．また，実験結果より提案手法の有効性が証明できた．動静判定の閾値を動的に決定しながら，このシステムのリアルタイム性を高めることが今後の課題である．

文　　献

[1] 村井和昌，中村哲，“ 画像と音声情報の併用による雑音に頑強な発話検出，”情報処理学会研究報告，音声言語情報処理 37-10 pp.55-60 (2001.7.14)

[2] 村井和昌，野間啓介，熊谷建一，松井知子，中村哲，“ 口周囲画像による頑強な発話検出，”情報処理学会研究報告，音声言語情報処理 37-13 pp.73-78 (2000.12.21)

[3] 川戸慎二郎，内海章，桑原和宏，“ あなたの顔をインターフェースに実時間処理で目，鼻，口を入力デバイスに使う，”画像の認識・理解シンポジウム (MIRU2004)，pp.I-107-108，(2004.7)

[4] 浅野英輔，荻原昭夫，柴田浩，“ 不特定話者に対する唇形状抽出法，”信学論，A Vol.J85-A No.3 pp.3998-402 (2002.3)

[5] 中田康之，安藤護俊，“ 色抽出法と固有空間法を用いた読唇処理，”信学論，D-II Vol.J85-D-II No.12 pp.1813-1822 (2002.12)

[6] 吉永智明，田村哲嗣，岩野公司，古井貞熙，“ 横顔画像から抽出した口唇角度情報を用いたマルチモーダル音声認識，”日本音響学会 2004 年秋季研究発表会講演論文集，vol.1，3-1-19，pp.147-148 (2004-9)．

[7] 泉正夫，藤本健雄，福永邦雄，“ SNAKE モデルを用いたエッジ線画像抽出とその物体認識への応用，”信学論，D-II Vol.J75-D-II No.12 pp.2010-2017 （1992.12)

[8] 菊池稔，木村元，小林重信，“ 遺伝的アルゴリズムを用いた赤外線画像からの移動目標検出システム，”信学論，D-II Vol.J84-D-II No.10 pp.2224-2233 (2001.10)

[9] 横川勇仁，船曳信生，東野輝夫，小田政志，森悦秀，“ Deformable Tamplate マッチング法による唇輪郭抽出法の改良と歯科医療応用を目的とした評価，”信学論，D-II Vol.J86-D-II No.8 pp.1177-1185 (2003.8)

[10] 関岡哲也，横川勇仁，船曳信生，東野輝夫，山田朋弘，森悦秀，“ 関数合成による唇輪郭抽出法の提案，”信学論，D-II Vol.J84-D-II No.3 pp.459-470 (2001.3)

[11] 須賀弘道，羽鳥好律，樽松明，“ SNAKE を用いた顔画像からの構成部品の輪郭抽出，”信学論，A Vol.J79-A No.2 pp.298-301 (1996.2)

[12] 柏野邦夫，黒住隆行，村瀬洋，“ ヒストグラム特徴を用いた音や映像の高速 AND/OR 探索，”信学論，Vol.J83-D-II，No.12，pp.2735-2744(2000.12)

[13] 柏野邦夫，ガビン　スミス，村瀬洋，“ ヒストグラム特徴を用いた音響信号の高速探索法-時系列アクティブ探索法-，”信学論，D-II，Vol.J82-D-II No.9 pp.1365-1373 (1999.9)

[14] 木村昭悟，柏野邦夫，黒住隆行，村瀬洋，“ グローバルな枝刈り

を導入した音や映像の高速探索，”信学論，D-II Vol.J85-D-II No.10 pp1552-1562 (2002.10)
[15] Robert E.Schapire, Yoram Singer,“ Improved Boostiong Alogirithm Using Confidence-rated Predictions, ”*Machine Learning*, 37(3)：297-336, 1999
[16] Norbert Binder, Konstantin Markov, Rainer Gruhn, Satoshi Nakamura: “ SPEECH NON-SPEECH SEPARATION WITH GMMS ”, 日本音響学会講演論文集 2001 年 10 月, pp141-142.
[17] DivX.Inc,“ http://www.divx.com/ ”